Comparative Analysis of Three Typical Text Mining Software

Xiaguang You
Electronic Information Department
Zhengzhou Electric Power College
Zhengzhou 450000, China
youxiaguang@sina.com

***Abstract*---- Based on brief introduction of the functions of three text mining software named as SAS Text Miner, VisualText, and TRS CKM, this paper comparatively analyses the three text mining software from three aspects of features embedded in data preparation, data analysis and result reporting. It is found that these software have their own characteristics and strengths.**

Keywords- text mining; SAS Text Mining； VisualText； TRS CKM

People urgently need convenient and practical tools to extract knowledge what they need from large-scale heterogeneous information resources because of the massive popularity of the Internet, the notable improvement of enterprises' informationization and the increasing promotion of the significance of knowledge. Hence various kinds of texting mining tools are developed such as ext Analyst, MegaSearch, PolyAnalyst, ICrossReader, Yahoo Planet, Dataset, ThemeScape, IpServer, Taxonomy engine, TextFinder, ZyIMAGE, Knowledge Server, In Query, Texis, SearchExpress, Visual Text, Textpack, SAS Text Miner, TRS CKM, WordStat, SPSS Clementine and so on among which SAS Text Miner, VisualText, and TRS CKM are three typical text mining software. Based on brief introduction to these three software, this paper intend to compare and analyze their characteristics and functions so that to make people develop or use text mining tools to do knowledge finding and management in a better way.

I. BRIEF INTRODUCTION TO THREE TYPICAL TEXT MINING SOFTWARE

A. SAS Text Miner

SAS is a business analytics software and services provider who is globally leading. SAS Text Miner is an additional product of SAS Enterprise Miner. It's a useful tool which can find and extract knowledge from text files. It transforms text data into usable and comprehensible format which can promote classify files, find specific relationship between files or correlate and cluster files.

1) Function: SAS Text Miner includes the following functions [1]:

- Accessing to common data. This includes the following: Access to various kinds of text data such as PDF, extended ASCII text, HTML and word; the ability of web crawling; the ability of extract, transform and load text data into SAS data set which is provided for text mining.
- Supporting multiple languages. This includes the following: English, French, German, Italian, Portuguese, Spanish, Chinese (traditional and simplified); supporting Latin-1, double byte character set and UTF-8 code.
- Owning self-document interface. This includes the following: Eliminating manual keyboard input by visual figures; modifying, saving and sharing procedure flow chart; allowing simplified-HTML format to come out various data.
- The ability to preprocess with broad text. This includes the following: Obtaining and refining the most potential information in corpus; automatic spell check; labelling the property of a word based on sentence situation; extracting noun group; users can customize and preinstall the list of synonyms; dividing compound words into different subclass words.
- Doing extensive characteristics extraction. This includes the following: Customized dictionary can extract specific information such as people's name, products, organizations URLs and addresses; normalizing extracted substances and listing them into matrix tables.
- The technology of dimensionality reduction. This includes the following: Preprocessing text data into information matrix; identify files' n highest weighted words automatically; transform every file into an n-dimension subspace by using singular value decomposition.
- Text clustering. This includes the following: Clustering files based on files content, do highest anticipant clustering by space clustering technology; doing automatically classifying to files and form classification system by polymerization method's layering clustering based on Ward polymerization method; doing files clustering according to procedure flow chart; using primitive files'

additional structural data such as age, habits and so on to do document clustering.

2) Operating Environment: SAS Text Miner can support several platforms such as AIX, HP-UX Itanium, Linux, Windows, Solaris on SPARC and so on; it can also support several browsers such as IE and Firefox; users need to install two software Sun JRE and SAS Enterprise Miner.

B. VisualText

VisualText is an ideal tool developed by Text Analysis International, Inc to do quick accurate information extraction, natural language processing and txt analyzing. It not only can accurately analyze extracted recovering information and also can be a system classifying web, a analyzer monitoring financing chatting, an email analyzer and a selectable web spider and so on.

1) Characteristics and advantages: The characteristics and advantages of VisualText are listed in table 1[2].

TABLE I. THE CHARACTERISTICS AND ADVANTAGES OF VISUALTEXT

Characteristics	***Advantages***
composite developing environment	reducing resources that developing text analyzer needs
enriching composite graph user interface tool and data view	speeding up the developing of senior and accurate analyzer
NLP++ programming language	anything imagined can be programmed; users dedicate into tasks but not details
scalar knowledge base management system	easy to construct noumenon, dictionary, semantics and other demand models
automatic rule generation	preserve and improve analyzer easily by emphasizing text
NLP++ comprehensive programming, ruling, parse tree and knowledge base	easy to visit and modify environment
compiling of analyzer and knowledge base	fast execution operating environment
debugging analyzer's optimal graph user interface support	reducing analyzer's debugging time and work load
multichannel analyzer	modularized, maintainable, readable and extendible
single parse tree	can generate effective analyzers
independent normal form; synthesizing several forms and styles of text analyzing	helpful for developer to use the most suitable method for handy tasks
adding the user project of C++ programming	easy to synthesize third party's software's open structure; extending user's central ability
specific situation's rules of operation	higher accuracy and faster implementing
built-in trust mechanism	easy to calculate and evaluate trust

2) Functions and applications: VisualText has the following functions and applications:

- Text extraction: extracting and connecting important information from texts and standardize them; it can also recognize text's name, location, date, and other fine features accurately.
- Text indexing: using indexing function to support the ability to retrieve high-quality text World Wide Web and other electronic sources.
- Text filtering: can determine whether a document is relevant accurately and quickly.
- Data mining: can find important information among a lot of texts.
- Test rating: can do reading and rating to random testing by using ideal answers.
- Text abstracting: can form brief and accurate description of the text content.
- Automatic coding: can do recovery processing and automatically form related reports.

3) Operating Environment: Text Analyst can be applied in Windows.

C. TRS CKM

TRS CKM a domestically leading Chinese text mining software that is developed by TRS. TRS CKM integrates a number of TRS Chinese information processing technology that can provide strong development interface for Chinese text mining applications. It can be applied in enterprise knowledge portals, value-added information services, intelligent search engines, digital libraries, intelligence analysis, information security and filtering, e-commerce systems and so on.

1) Functions: TRS CKM has the following main functions[3]:

- Text classification: Improving classification accuracy and speed of classification by using content-based automatic text classification, rule-based automatic text classification technology and feedback learning mechanisms and supplementary training mechanisms; comprehensive content classification, rule classification and two or more classification techniques; supporting high-accuracy multi-level classification and the classification of hybrid of Chinese and foreign languages.
- Natural language searching: Supporting text similarity search and natural language search by using text-based documents "fingerprint" re-check technology and advanced natural language processing techniques; doing content correlation calculation according to the location and frequency and other parameters of the key words in the content; outputting search results according to high to low content relevance.
- Text clustering: providing massive documents' visual analysis and integrated applications by using

automatic clustering technique which is based on similarity algorithm.

- Text abstracting: extracting article main topics automatically by using text automatic abstract techniques which is based on statistics.
- Text filtering: effectively identifying and filtering harmful or garbage text messages by using automat text filtering technology which is based on statistics and machine learning to help users get rid of the intrusion of harmful information.
- Pinyin searching: providing users with homophones query suggestions by using phonetic technology and multi-tone raw technology which is based on statistics to help users search more effectively.
- Related phrases searching: obtaining relevant phrases and to enhance search performance by using related phrases technology which is based on artificial soring and data mining combines.
- Common sense proofreading: effectively identifying politically sensitive text error message by using semantic-based proofreading technology to avoid accidents and adverse effects of political propaganda.
- Text word segmentation: cutting Chinese characters sequence into meaningful words by using the combination of rule-based and statistical segmentation techniques to improve searching relevance ranking.

2) Operating Environment: TRS CKM can be applied in Windows and Linux.

II. THE COMPARISION OF THREE TEXT MINING SOFTWARE

To further illustrate the differences among the three software we can compare from three aspects of the text mining: data preparation, data analysis, functions in result reports[4] which show in the Table 2.

III. CONCLUSION

Just like the above shows, although there are some similarities in function among three text mining software, such as they all emphasize "universal data access", "text analytical and extraction", "text clustering", "text summaries," "natural language query", "interactive window result" and "multilingual support" there still exist obvious differences between them. For example, only SAS Text Miner and TRS CKM have the functions of "word segmentation" and "text similarity search"; only VisualText owns the functions of "Indexing", "automatic coding" and "visualize results show"; only SAS Text Miner owns the function of "dimensionality reduction technology"; As a domestically leading text mining software, though lacks some important functions compared to foreign similar software, TRS CKM still has its owning characteristics such as "phonetic searches", "related phrases searches" and "common sense proofreading" which other software don't have. And domestic and foreign text mining software are also different from the operating environment. So this comparative analysis can provide helpful useful reference to users when then choose proper text mining software or develop more powerful new software.

REFERENCES

[1] Text mining with SAS® text miner[EB/OL].[2010-1-26] http://www.sas.com/technologies/analytics/datamining/textminer/index.html

[2] VisualText features and benefits[EB/OL]. [2010-1-28] http://www.textanalysis.com/intro.htm

[3] TRS text mining software （TRS CKM） [EB/OL].[2010-1-28]. http://www.trs.com.cn/products/textmine/trsckm/

[4] Segall R S, Zhang Q. A survey of selected software technologies for text mining [M]//Song M, Wu Y B. Handbook of research on text and web mining technologies. Hershey: Information Science Reference, 2008: 776-784

TABLE II THE COMPARISON OF THREE TEXT MINING SOFTWARE

Functions \ ***Software***		***SAS Text Miner***	***Visual Text***	***TRS CKM***
Data preparation	Universal data access	√	√	√
	Text analysis and extraction	√	√	√
	Custom dictionary	√	√	
	Automatic text cleaning	√	√	
Data analysis	Classification	√	√	√
	POS tagging	√		√
	Filtering		√	√
	Concept linking	√	√	
	Word segmentation	√		√
	Text clustering	√	√	√
	Feature extraction	√		√
	Text abstraction	√	√	√
	Automatic coding		√	
	Dimensionality reduction technique	√		
	Natural language query		√	√
	Indexing		√	
	Text similarity search	√		√
	Pinyin search			√
	Common sense proofreading			√
Result report	Interactive window results	√	√	√
	Visualization results show		√	
	Vocabularies sorting	√		
	Multilingual support	√	√	√